Vaccine (2008) **26**, 2951—2965

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/vaccine

Detecting AIDS restriction genes: From candidate genes to genome-wide association discovery

H.B. Hutcheson[a], J.A. Lautenberger[a], G.W. Nelson[b], J.U. Pontius[b], B.D. Kessing[b], C.A. Winkler[b], M.W. Smith[b], R. Johnson[b], R. Stephens[c], J. Phair[d], J.J. Goedert[e], S. Donfield[f], S.J. O'Brien[a,*]

[a] *Laboratory of Genomic Diversity, National Cancer Institute-Frederick, Frederick, MD 21702, USA*
[b] *Laboratory of Genomic Diversity Basic Research Program, SAIC- Frederick, Inc., NCI-Frederick, Frederick, MD 21702, USA*
[c] *Advanced Biomedical Computer Center, SAIC-Frederick, Frederick, MD 21702, USA*
[d] *Northwestern University Medical School, Division of Infectious Diseases, Chicago, IL 60611, USA*
[e] *Viral Epidemiology Branch, National Cancer Institute, Rockville, MD 20852, USA*
[f] *Rho, Inc., Chapel Hill, NC 27514, USA*

Received 29 October 2007; received in revised form 18 December 2007; accepted 19 December 2007
Available online 1 February 2008

KEYWORD
AIDS restriction genes (ARGs)

Summary The screening of common genetic polymorphisms among candidate genes for AIDS pathology in HIV exposed cohort populations has led to the description of 20 *A*IDS *r*estriction *g*enes (ARGs), variants that affect susceptibility to HIV infection or to AIDS progression. The combination of high-throughput genotyping platforms and the recent HapMap annotation of some 3 million human SNP variants has been developed for and applied to gene discovery in complex and multi-factorial diseases. Here, we explore novel computational approaches to ARG discovery which consider interacting analytical models, various genetic influences, and SNP-haplotype/LD structure in AIDS cohort populations to determine if these ARGs could have been discovered using an unbiased genome-wide association approach. The procedures were evaluated by tracking the performance of haplotypes and SNPs within ARG regions to detect genetic association in the same AIDS cohort populations in which the ARGs were originally discovered. The methodology captures the signals of multiple non-independent AIDS-genetic association tests of different disease stages and uses association signal strength (odds ratio or relative hazard), statistical significance (*p*-values), gene influence, internal replication, and haplotype structure together as a multi-facetted approach to identifying important genetic associations within a deluge of genotyping/test data. The complementary approaches perform rather well and predict the detection of a variety of undiscovered ARGs that affect different stages of HIV/AIDS pathogenesis using genome-wide association analyses.
Published by Elsevier Ltd.

* Corresponding author. Tel.: +1 301 846 1296; fax: +1 301 846 1686.
E-mail address: obrien@ncifcrf.gov (S.J. O'Brien).

0264-410X/$ — see front matter. Published by Elsevier Ltd.
doi:10.1016/j.vaccine.2007.12.054

The discovery that HIV entered cells by binding first to CD4 then to CCR5 was pronounced in simultaneous articles by five outstanding research groups in 1996 in the pages of *Science*, *Nature*, and *Cell* [1—6]. This seminal announcement led immediately to the discovery of *CCR5-*Δ*32*, the first human ARG by which homozygous carriers were near completely resistant to HIV-1 infection, regardless of the extent of exposure [7—9]. Since then researchers in the NCI's Laboratory of Genomic Diversity have used genetic associations studies investigating candidate genes with assembled HIV/AIDS cohort populations (~10,000 study participants) to describe some 20 *A*IDS *r*estriction *g*enes (ARGs) that involve HIV entry, innate or acquired immunity, and HIV transcriptional regulation (Table 1) [10—13]. Demonstrated genetic resistance by human variants in HIV entry receptors has led to the birth of a new generation of anti-HIV therapy, termed HIV entry inhibitors, including fuzeon-T20 maraviroc, (approved for salvage AIDS therapy by US-FDA), and several compounds now in clinical trials [14—20].

The ARG discoveries have become a harbinger for genetic association studies in other complex genetic diseases including cancers, infectious disease such as hepatitis B and C, malaria, and chronic degenerative diseases. Yet all the ARGs discovered to date involve the candidate gene approach whereby advances in virology, immunology, structural biology, or model systems have pinpointed potential loci that collaborate with HIV in pathogenesis. Further, the known ARGs account for approximately 10% of the epidemiological variance that characterizes AIDS pathogenesis [10,21]. This means there are 10 times more undiscovered influences for the dynamics of AIDS yet to be discerned than the known ARGs can explain.

The Human Genome Project provided a draft sequence initially in 2001 and a more polished completed version in 2003 [22]. Included in the human genome annotation has been the assessment of some 9 million single nucleotide polymorphisms (SNPs) and their linkage disequilibrium (LD) based non-random association in the 2006 release of the NHGRI-funded HapMap project (Phase II) [23,24]. The combination of high-throughput genotyping platforms and the recent HapMap annotation of some 3 million human SNP variants have been developed for and applied to gene discovery in complex and multi-factorial diseases. Varying opinions have emerged within the human genetic literatures as to the ideal strategy for genome-wide association (GWA) in complex multi-factorial diseases such as AIDS [25—33]. A particular challenge is the avoidance of false positive disease association signals that can arise due to statistical fluctuations that fail to replicate and can mislead the field [34—37]. As a prelude to the transition from candidate gene detection (Table 1) to GWA based ARG discovery, we explore some of these issues empirically with assembled AIDS cohorts and known ARGs.

Genome-wide association prospects

A major challenge for genome-wide genetic association studies involves the efficiency of linkage disequilibrium (LD) with adjacent ''proxy'' SNPs to identify disease gene causal

Table 1 Human AIDS restriction gene (ARGs) that affect HIV-1 infection, AIDS progression, and AIDS outcome

	Gene	Allele	Mode	Effect	Time
(1)	*CCR5*	Δ*32*	Recessive	Prevent infection	—
	CCR5	Δ*32*	Dominant	Prevent lymphoma	Late
	CCR5	Δ*32*	Dominant	Delay AIDS	Overall
(2)	*CCR5P*	*P1*	Recessive	Accelerate AIDS	Early
(3)	*CCR2*	*64I*	Dominant	Delay AIDS	Overall
(4)	*SDF1*	*3′A*	Recessive	Delay AIDS	Late
(5)	*EOTAXIN-MCP1*	*Hap7*	Dominant	Enhance infection	—
(6)	*RANTES*	*−403A*	Dominant	Accelerate AIDS	Overall
		In1.1C	Co-dominant	Accelerate AIDS	Overall
(7)	*HLA*	*A,B,C, ''Homozy''*	Co-dominant	Accelerate AIDS	Overall
(8)	*HLA*	*B*35Px*	Co-dominant	Accelerate AIDS	Overall
(9)	*HLA*	*B*57*	Co-dominant	Delay AIDS	Overall
(10)	*HLA*	*B27*	Co-dominant	Delay AIDS	Overall
(11)	*KIR*	*3DS1*	Epistatic (Bw4-801)	Delay AIDS	Overall
(12)	*IFNG*	*179T*	Dominant	Accelerate AIDS	Overall
(13)	*IL10*	*5′A*	Dominant	Limit infection	—
	IL10	*5′A*	Dominant	Accelerate AIDS	Late
(14)	*CXCR6*	*E3K*	Dominant	Accelerate PCP	Late
(15)	*APOBEC3G*	*H186R*	Recessive	Accelerate AIDS	Overall
(16)	*TSG101*	*Hap2*	Dominant	Accelerate AIDS	Early
(17)	*DCSIGN*	*−336T*	Dominant	Decrease infection	—
(18)	*TRIM5*	*Hap4*	Dominant	Increase infection	—
(19)	*Cul5*	*HapI*	Co-dominant	Accelerate CD4 loss	—
(20)	*PP1A* (cylophilinA)	*SNP-4*	Dominant	Accelerate AIDS	—

Primary citations in [10—12,47—52].

or operative oSNPs; that is, to track and detect genetic influence above the background of statistical fluctuations necessarily associated with the large numbers of association tests (oSNP is the operative/causal SNP or indel variant that confers resistance/susceptibility to HIV/AIDS). The difficulty is emphasized by genetic association studies that fail to replicate due to low case numbers, low frequencies of oSNPs, low relative risk of the oSNP-bearing genotypes for the disease, and with mis-identification of the oSNP versus the proxy-p SNPs [10,33—37]. Further, many GWA studies initially discount very significant associations that do not achieve ''Bonferoni correction'' *p*-values or that of the most extreme hits, perhaps missing actual genetic influences in a sea of false positives [38—42]. Although informative theoretical and simulation approaches to these issues have appeared, an empirical test of the pitfalls and strengths of GWA would be illuminating. To accomplish such an experiment we examined how well adjacent SNPs, multi-SNP-haplotypes in the region, and a well-characterized study population (cohorts used to implicate the original ARGs) would enable determination of a true genetic association if the oSNP had been unknown.

We designed a ''pilot study'' where 306 SNPs spaced at 15—18 kb density across the regions of eight previously validated ARGs, (Table 2), were genotyped and tested for association with different stages of HIV/AIDS disease. Certain ARGs have few neighboring genes (*IL10-5A*, *SDF1*), while others are nested within gene clusters (*CCR2-64I-CCR5-P1-CCR5-Δ32*; *EOTAXIN-MCP1-MCP2*; Fig. 1. SNP genotypes were assessed among 2139 patients at risk for AIDS from the epidemiological study cohorts originally used to discover the ARGs [10—12]. Pair-wise LD was determined, haplotype blocks were delineated, and haplotypes were defined by their included SNP alleles. Our goal was to explore and attempt to answer the following questions: (1) How well and how frequently do we track the oSNP with one or more demonstrated pSNP variants on haplotypes in strong LD with the causal oSNPs, and how often would we miss it? (2) Given a haplotype structure of a given candidate gene region, do haplotype associations improve chances for oSNP detection? (3) Can we develop adequate computational routines that facilitate inspection and interpretation of very large numbers of genetic association tests? (4) What are the implications of these empirical association tests for feasibility and strategy of GWA studies for AIDS or for other complex diseases?

Detecting known ARGs using close adjacent SNPs

A group of 306 SNPs flanking each of seven ARGs on five chromosomes (Table 2) plus a region selected as a negative control for AIDS (chromosome 7q36 containing *CFTR* gene) were genotyped in 2139 particularly informative European American study participants (based on clinical assessment of AIDS progression, see Supplemental Methods) using an Illumina (243 SNPs) or Sequenom (92 SNPs) genotyping platforms (Table 2, Supplemental Table 1). The average density is 1 SNP/17 kb with block sizes, number of blocks, number of haplotypes, mean haplotype size, and range for each region listed in Table 2. Fig. 1 shows ARGBROWSER,

Table 2 Patterns of SNP and haplotype variation in six regions Including 8 ARG variants and a control *CFTR* region[a]

Chr region	ARG-oSNP[b]	Map coordinates		Length (kb)	No. SNPs	SNP density (kb)	Haplotypes[b]			
							Number		Hap block length (kb)	
		pter	qter				Blocks	Haps	Mean	Range
1q31-32	*IL10-5′A*	203,851,217	204,384,080	533	33	16.1	13	106	75.6	11.1—146.6
3p21-22	*CCR5/2 (Δ32; P1; 64I)*	45,548,733	46,663,063	1,115	68	16.4	34	262	102.8	11.9—189.9
10q11	*SDF1-3′A*	43,642,225	44,598,838	956	49	19.5	19	162	118.4	29.6—262.7
17q12E	*EOTAXIN-Hap7*	32,387,585	32,967,726	584	40	14.6	13	88	56.4	25.2—112.9
17q12R	*RANTES-409:In1.1c*	32,976,593	34,374,495	1,402	72	19.5	28	214	107.6	5.5—318.2
7q36	*CFTR*[a]	116,639,477	117,410,794	771	44	17.5	21	204	121.78	70.0—202.6
			SUM:	5,361	306	Avg. 17.3	128	1010	Avg. 101.8	Overall: 5.5—318.2

[a] The *CFTR* region on chromosome 7 serves as a negative control for the ARG discovery since this region, though well studied, has no anticipated or demonstrated influence on HIV/AIDS.
[b] For details of map region and haplotype construction, see Supplemental Methods and Supplemental Fig. 1, Supplemental Tables 2 and 3.

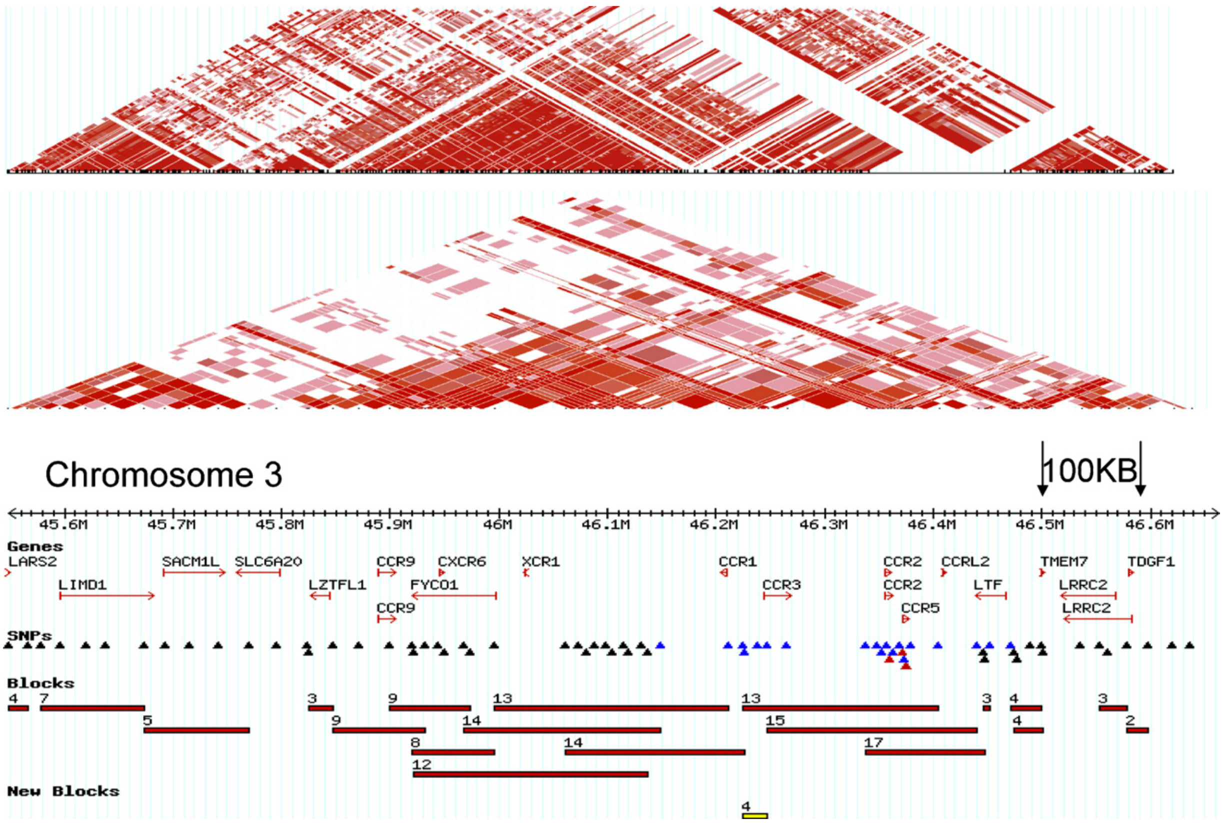

Figure 1 ARGBROWSER for *CCR2/5* chromosome 3 region. Top-LD from Haploview; bottom from BLOCKHEAD; see Supplemental Methods. Gene coordinates follow UCSC human genome browser. Red triangles — oSNPs; blue triangles — pSNPS in LD with oSNP; black triangles — adjacent SNPs typed across the region. Red bars — haplotype blocks with number of included haplotypes listed above (see Supplemental Tables 2 and 3) Yellow bars are ''NEW'' blocks that appear from population genotypes when oSNP is not considered in haplotype construction.

a generic genome browser display of physical genetic map of the *CCR2/5* chromosome-3 region, the included SNPs genotyped, pair-wise linkage disequilibrium detected, and the discerned overlapping haplotype block structure of the SNPs as described in the Supplemental Methods. The SNP genotypes were used to define haplotypes and haplotype blocks within each ARG region ($N = 1010$ haplotypes; Table 2, Supplemental Tables 2 and 3). Then haplotypes were rebuilt excluding the oSNP from the data to produce a second group of haplotypes ($N = 997$ haplotypes without the oSNP; Supplemental Tables 2, 3 and 7).

A total of 136 non-independent statistical association tests were designed to reveal genetic influences of previously published and validated ARGs (Table 3; Supplemental Tables 4 and 5) with various AIDS outcomes. The tests reflect four stages of HIV/AIDS pathogenesis: (1) HIV-1 infection, 12 tests; (2) AIDS progression using categorical groups, e.g. fast versus slow progressors to AIDS post-HIV infection, 28 tests; (3) survival analysis — 72 tests; and (4) AIDS defining disease or sequelae — 24 tests. Every SNP (including the oSNP and designated pSNP) plus every haplotype were tested for each of the 136 AIDS association tests. In all we determined 654,534 genotypes, and performed 41,616 SNP tests, 137,360 Hap tests (+oSNP), and 135,592 Hap tests (for discerned haplotypes after oSNP was removed), a total of 314,568 genetic association tests. Previously published hazard/odds ratios, *p*-values, number of study participants, and citation for each oSNP reported to implicate each ARG, plus the same values for the present study population are presented in Supplemental Table 6 as a starting point for assessment of pSNP and haplotype analyses.

Computational tools for ARG discovery

Two new computational approaches, ARGARRAY and ARGRANK were developed to identify the genetic associations from GWA studies. ARGARRAY visually displays the SNP-genetic association signal strength (*p*-value) for the 136 ARG tests in a horizontal line of squares where the color (heat plot) discriminates the statistically strong associations from the weaker and non-significant effects (Fig. 2). To examine a genomic region, the horizontal heat plots for adjacent SNPs are aligned in the same order as the SNP markers occur on the chromosome. Therefore, all the adjacent markers irrespective of their LD relationship can be inspected together in a two-dimensional color matrix that captures 136 AIDS association tests (horizontal axis) and each SNPs or haplotype (vertical axis). Clusters of highly significant genetic associations (beige — $p > 0.05$; yel-

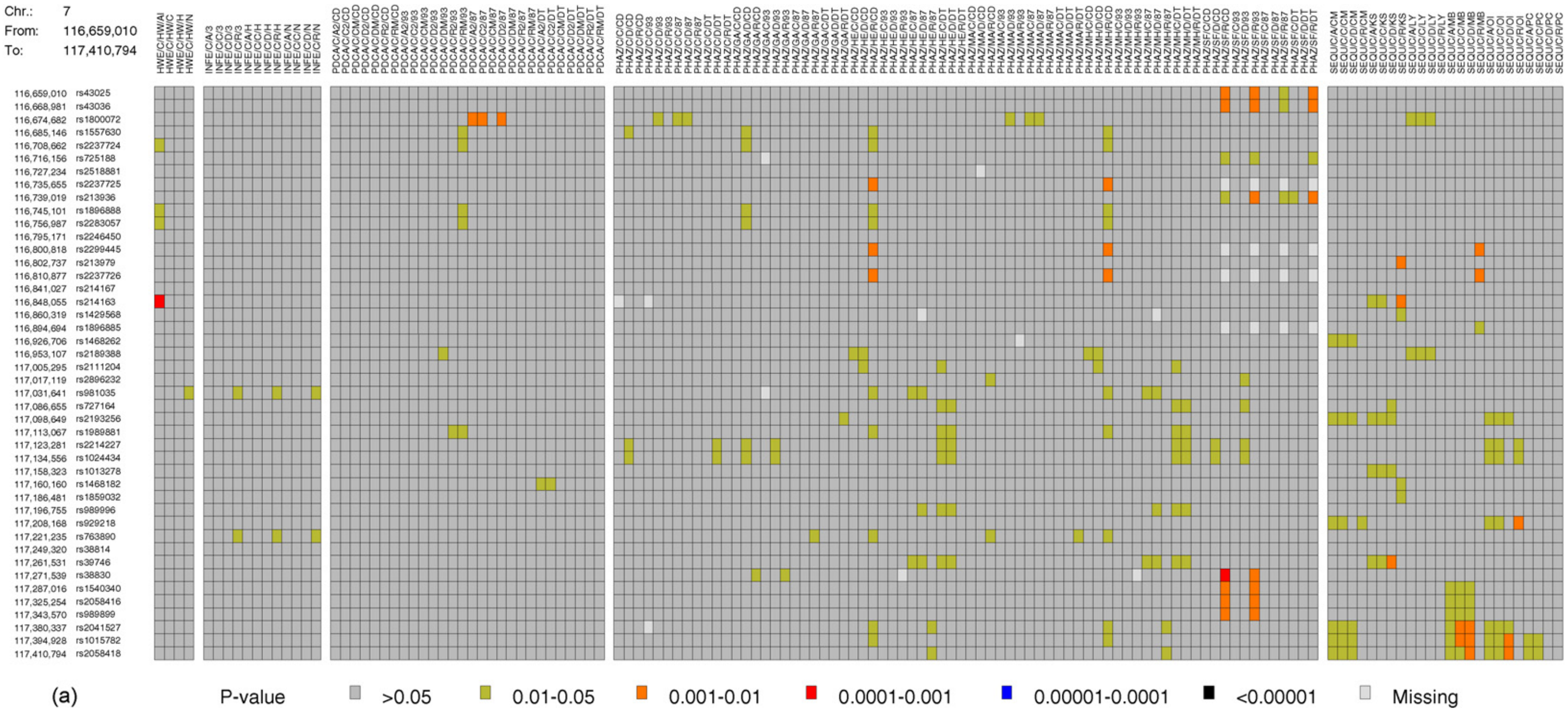

Figure 2 (a) ARGARRAY is a computational toll for visualization of the *p*-values for multiple non-independent genetic association tests (Table 3) for each SNP as a color ''heat'' plot compared to adjacent SNPs across tested gene region. This display captures replicated association signals derived from multiple test associations as well as multiple proxy SNPs in linkage disequilibrium with the oSNP (see text). Here, we depict ARGARRAY for 136 AIDS association tests (top, see Table 3 and Supplemental Tables 4 and 5) assessed for 44 SNPs (left) spaced at 17 kb across the ''negative control'' CFTR gene region on chromosome 7 (5984 SNP-test combinations). We also add 4 tests for Hardy—Weinberg Equilibrium (HWE) on the left. A physical map of the SNPs, Haps, LD and map coordinates for the chromosome 7 CFTR region is presented in Supplemental Fig. 1a. Color key indicates significant *p*-values of increasing significance; (b) ARGARRAY for chromosome 1-*IL10* region; oSNP names on *Y*-axis are red; pSNPs are blue; (c) ARGARRAY for chromosome 3, including *CCR5-*Δ*32*, *CCR5-P1* and *CCR2-64I*. Colors of SNPs as in (b) plus green for SNPs with $D' > 0.8$ with oSNPs, but located outside haplotype blocks defined in Supplemental Fig. 1c, and Supplemental Tables 2 and 3.

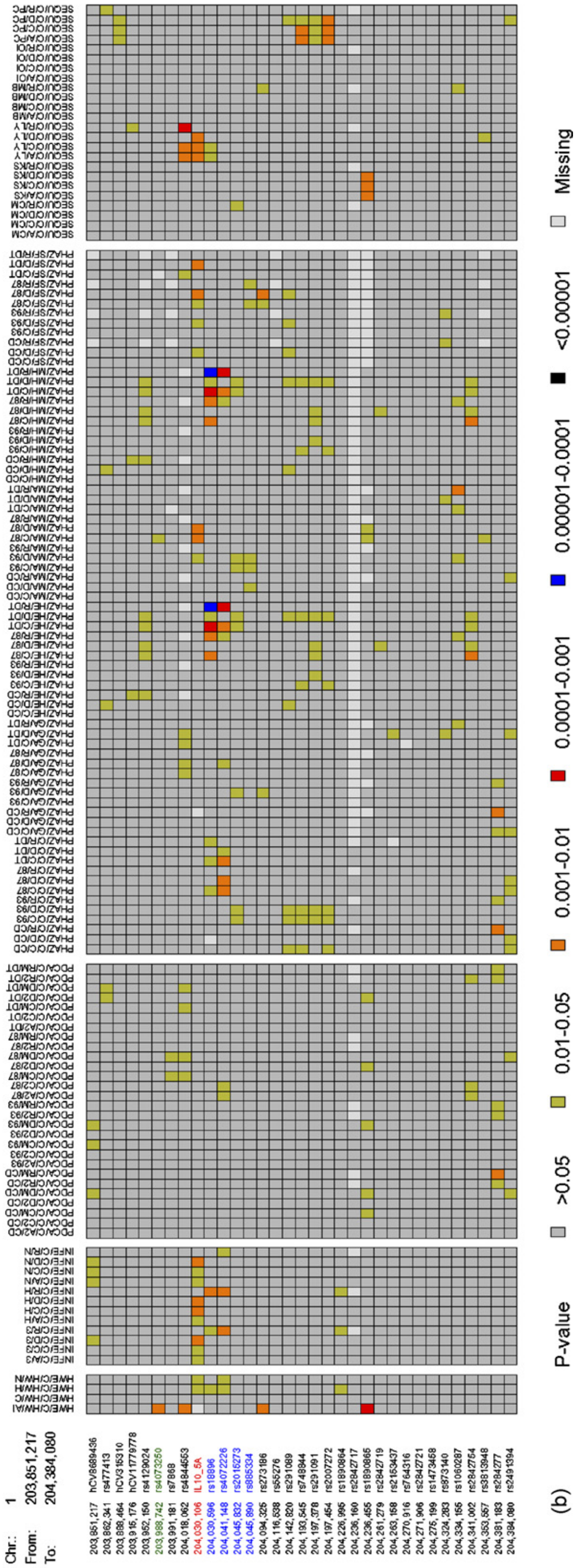

Figure 2 *(Continued)*

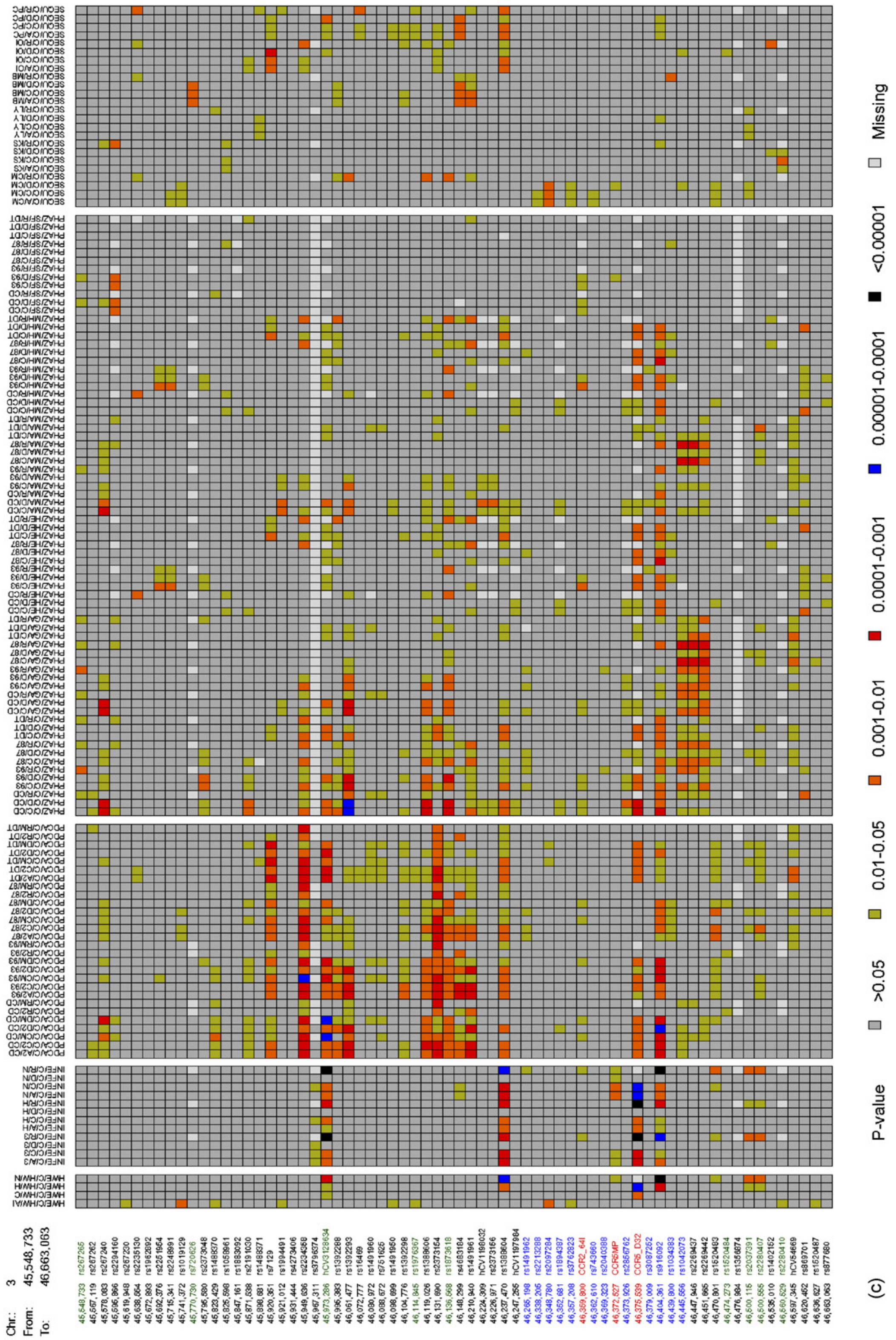

Figure 2 *(Continued)*.

Table 3 List of 136 genetic association tests used to input ARGARRAY and ARG RANK

Genetic hypothesis	Number of tests	Variables in each test
I. Infection	12 tests	3 comparisons (SN vs. SC, HREU vs. SC, HREU vs. SN vs. SC) × 4 modes (allelic, codominant, dominant, recessive)
II. Progression—categorical	28 tests	4 outcomes (CD4 < 200, AIDS-1993, AIDS 1987, death) × 7 modes (allelic—dichotomous, codominant—dichotomous, codominant—multipoint, dominant—dichotomous, dominant—multipoint, recessive—dichotomous, recessive—multipoint).
III. Progression—survival	72 tests	4 outcomes (CD4 < 200, AIDS-1993, AIDS-1987, death) × 3 modes (codominant, dominant, recessive) × 6 populations (Euro. Amer., homosexuals, hemophiliacs, MACS, MHCS, SFCC)
IV. Sequelae	24 tests	6 AIDS-defining conditions (CMV, KS, lymphoma, Mycobacterial infection, OI, PCP) × 4 modes (allelic, codominant, dominant, recessive)
V. Hardy—Weinberg	4 tests	4 (All subjects; SC, SN, HREU)

See also Supplemental Tables 4 and 5.
Abbreviations SC — seroconvertor; SP — seroprevalents; HREU — high risk exposed uninfected; OI — opportunistic infection; PCP — pneumocystis carnii pneumonia; KS — Kaposi's sarcoma; CMV — cytomegalovirus. MACS, MHCS, SFCC AIDS cohorts see reference [10].

low < 0.05; orange — $p < 0.01$; red — $p < 0.001$; dark blue $p < 0.0001$; black $p < 0.00001$) for both association tests and LD SNPs are easily drawn to the eye for closer inspection.

ARGARRAY results for SNPs across the ARG regions are illustrated in Fig. 2, and tabulated in Table 4. The ''negative control'' region Chromosome 7-CFTR (Fig. 1a) shows a background pattern with 190 beige [$p < 0.05$] squares (~3.2% of 5984 test combinations) and 38 (0.6%) of the tests showing [$p < 0.01$] scores (Supplemental Table 8). This lower than expected incidence (we expect 5% and 1%, respectively) reflects the non-independence of the cumulative ARG association tests. The ARGARRAY for chromosome 1-*IL10* illustrates a positive result where both oSNPs and pSNPs show multiple [$p < 0.01$] signal squares for HIV infection, progression and sequelae tests (Fig. 2b). A more dramatic result came with chromosome-3 which contained three tightly linked ARGs (*CCR5-Δ32*, *CCR5-P1*, and *CCR2 64I*) plus a large backbone of linkage disequilibrium, resulting in 11—38 pSNPs that track the three ARGs (Table 4, Fig. 2c). The pSNPs include both those within the haplotype blocks (blue locus labels in Fig. 2b and c) as well as others outside the blocks but showing $D' > 0.8$ with the oSNPs (green in Fig. 2). The rich colors reflecting multiple highly significant tests and large LD across the region (Fig. 2c) are in dramatic contrast to the background of low color for the Chromosome 7 region (Fig. 2a). The complete ARGARRAY displays for SNPs and derivative haplotypes of each ARG region are presented in (http://home.ncifcrf.gov/ccr/lgd/) and the counts of [$p < 0.01$] are listed in Table 4 for the oSNPs and pSNPs.

A second computational tool, ARGRANK, plots five different rank values from the same 136 association tests (displayed in ARGARRAY) for each SNP or haplotype versus the position of the SNP on the map (Fig. 3). The algorithm consists of 5 rank criteria that assess strong genetic associations for each SNP (or haplotype) and compare these to other (SNPs) or haplotypes similarly assessed. The five ranking schemes capture significant *p*-values as well as relatively high odds/hazard ratios of a SNP compared to the other 305 SNPs in the screen (see Fig. 3 caption for rank criteria). In ARGRANK, a low score is desirable (reflected as a downward dip) as this reflects a higher ranking value.

The ARGRANK results (Fig. 3) tended to affirm the ascertainment of ARGARRAY. On Chromosome 1-*IL10* the oSNP and two adjacent pSNPs show consistent dips ($R < 50$) for five infection test ranks and for AIDS survival analyses rankings (Fig. 3a and b) in contrast to all the other SNPs across the *IL10* and other ARG regions. For the ARG-negative region, chromosome7-*CFTR*, two of the 44 SNPs ranked < 50 in test 1 (lowest *p*-value) and in test 3 (highest OR/p-val) for HIV infection, but not in the other infection ranks or in other genetic hypotheses (Fig. 3c and d). Absence of consistent dips across the five ranking schemes for two stages of HIV/AIDS (Fig. 3c and d) is illustrative of background statistical noise for ARGRANK. By contrast the chromosome-3 *CCR5/2* region showed multiple consistent low ranks (<50) for oSNP plus pSNPs, again reflecting the ARG signal and extensive LD in the region (Fig. 3e and f). Complete ARGRANK displays for SNPs, (and also for haplotypes with and without the oSNP included, see next section) for each ARG region are presented in http://home.ncifcrf.gov/ccr/lgd/.

To evaluate haplotype AIDS association detection, we determined haplotypes and haplotype blocks for each region (Supplemental Tables 2 and 3). The SNP allele structure for haplotypes that overlap the oSNP locus for each ARG is presented in Fig. 4 as well as the haplotype structure imputed across the same oSNP locus but after the oSNP was removed. In Table 5, we list haplotype blocks, their included haplotype frequencies, and an estimated ''*percent oSNP representation*'' (PSR) of a given oSNP-bearing haplotype. For example, if an oSNP is carried on two haplotypes with frequencies of 0.1 and 0.2, respectively in the population, the PSR of the first haplotype is 33.3% and the second 66.7%. Low PSR and further oSNP dilution in haplotypes where the oSNP is

Table 4 SNP scores of ARG-ARRAY and ARG-RANK in genetic association test

Chr. region	ARG-oSNP	ARG comput.	Genetic hypothesis—oSNPs					No. pSNPs in LD with oSNP[b]	Genetic hypothesis—pSNPs[b]					Total #SNP
		Tool	INFE (12)[a]	PRG-CA (28)	PRG-SA (72)	SEQ (24)	SUM		INFE (12)	PRG-CA (28)	PRG-SA (72)	SEQ (24)	SUM	
1q31-32	*IL10-5′A*	ARRAY ($p < 0.01$)	4	0	4	3	11	4	1	0	8	3	12	33
	rs 1800872	RANK ($1 < 50$)[c]	5	0	5	5	15		1	0	1	1	3	
3p21-22	*CCR5-Δ32*	ARRAY ($p < 0.01$)	9	13	18	0	40	38	22	93	134	22	271	68
	rs 333	RANK ($1 < 50$)	5	5	5	0	15		6	11	12	9	38	
3p21-22	*CCR5-P1*	ARRAY ($p < 0.01$)	2	0	0	0	2	11	9	13	21	3	46	68
	Rs 1799987	RANK ($1 < 50$)	5	0	0	0	5		1	1	1	1	4	
3p21-22	*CCR2-64I*	ARRAY ($p < 0.01$)	0	0	2	0	2	18	34	57	65	10	166	68
	Rs 1799864	RANK ($1 < 50$)	2	0	1	0	3		5	4	4	3	16	
10q11	*SDF1-3′A*	ARRAY ($p < 0.01$)	0	0	4	0	4	10	0	0	0	1	1	49
	Rs 1801157	RANK ($1 < 50$)[d]	2	2	5	0	9		0	0	3	4	7	
17q12-E	*EOTAXIN-HAP7*	ARRAY ($p < 0.01$)	0	0	0	0	0	14	3	2	2	0	7	40
	rs 4795895	RANK ($1 < 50$)[d]	2	0	0	0	2		1	1	2	0	4	
17q12-R	*RANTES-401*	ARRAY ($p < 0.01$)	0	0	0	0	0	4	0	0	0	0	0	72
	rs 2107538	RANK ($1 < 50$)[d]	0	0	0	0	0		0	0	0	0	0	
17q12-R	*RANTES-In1.1c*	ARRAY ($p < 0.01$)	0	0	0	0	0	8	0	0	0	2	0	72
	rs 2280789	RANK ($1 < 50$)[d]	0	0	0	0	0		0	0	0	1	0	
7q36·3-.7	*CFTR*[e]	ARRAY ($p < 0.01$)	–	–	–	–	0	44	0	3	22	13	38	44
		RANK ($1 < 50$)[d]	–	–	–	–	0		0	1	10	9	20	
Sum														306

[a] In parenthesis is number of genetic tests.

[b] pSNPs ($D' > 0.8$ with oSNP) are highlighted in blue in Fig. 1, and Supplemental Figs. 1–6.

[c] Left—(oSNP) list counts the number ARGRANK schemes (of the five listed in Fig. 3) that the oSNP ranks <50 relative to the other SNPs; Right-number of pSNPs identified for the specific ARG-oSNPs which rank <50 in 3 of 5 ranking schemes relative to the other 306 SNPs.

[d] See Supplemental Figs. 5b, 6b, 7b for ARGRANK plots of *SDF1*, *EOTAXIN*,*RANTES*, respectively.

[e] Counts for Chr 7 region include all 44 SNPs genotyped (5984 tests).

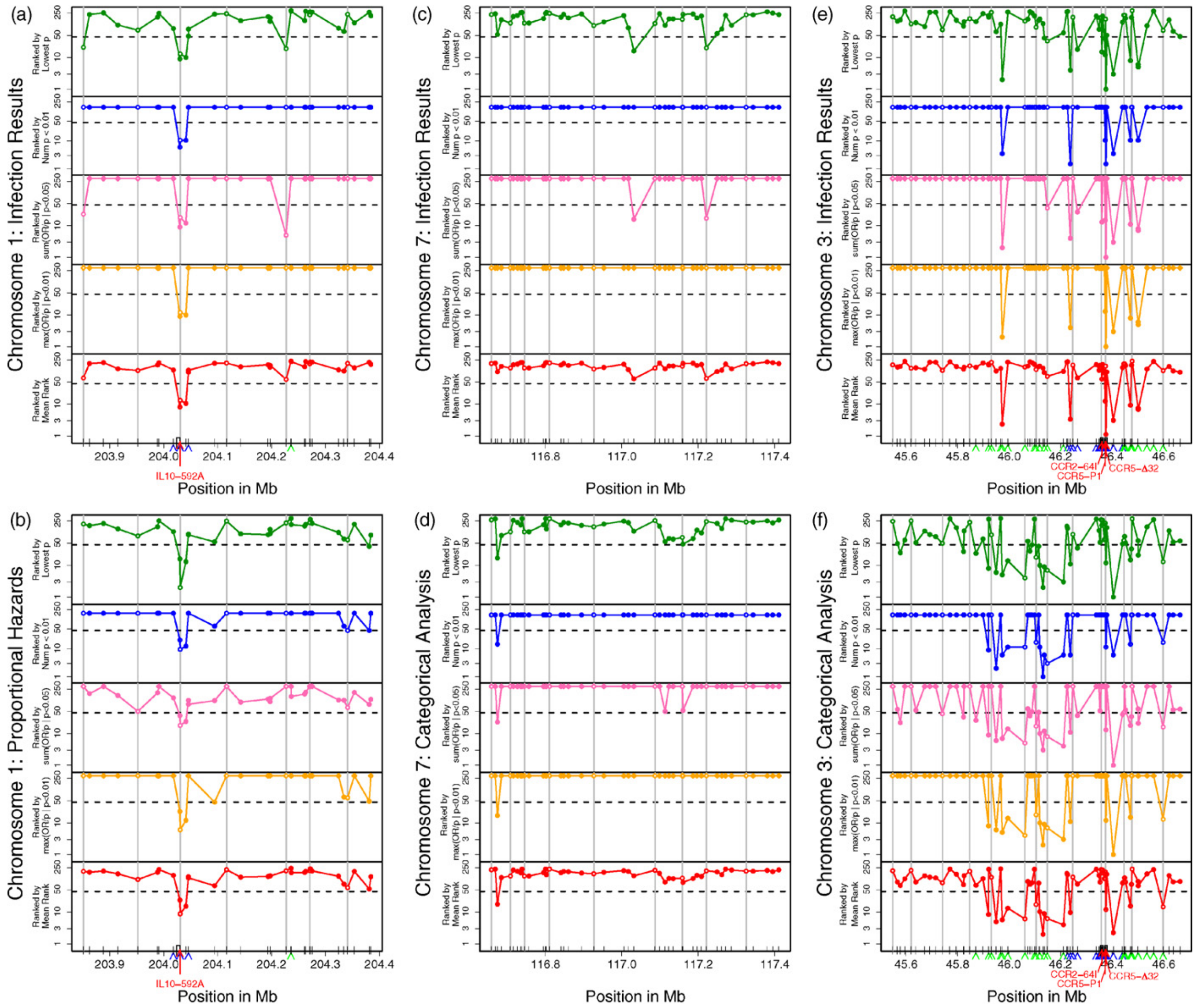

Figure 3 ARGRANK is a computational tool that compares extreme genetic association values (maximal odds ratios or minimum *p*-values) for a particular SNP across a group of analytical tests for a specific genetic hypothesis (e.g. for 12 HIV infection tests or 72 AIDS progression Survival-Cox Proportionate Hazards Tests, the same test used for ARGARRAY; see Table 3) to the same extreme values obtained for the other 305 SNPs in the study. The extreme values for each individual SNP are then ranked with respect to all other SNPs across the five ARG regions and each SNP's rank position is plotted versus its map coordinate position alongside the other SNPs in the region. Five different ranking criteria were computed and plotted for each SNP: (1) Rank of the lowest *p*-value (in 136 genetic association tests — Table 3; Supplemental Table 5) for a given SNP compared to the lowest *p*-value of the tests for other 305 SNPs; (2) Rank the number of tests where $p < 0.01$ for each SNP versus the number of tests ($p < 0.01$) for the other 305 SNPs; (3) Rank the sum of OR/*p*-value for tests with $p \leq 0.05$ for each SNP versus same for the other 305 SNPs; (4) Rank the maximum OR/*p*-value test with $p \leq 0.01$ for each SNP versus same for other 305 SNPs; and (5) Rank by the mean rank of a SNP in the previous four tests versus the mean rank of the same for the other 305 SNPs. (a) ARGRANK plots for HIV infection, chromosome 1-*IL10*, oSNP is red; (b) ARGRANK plots for HIV-AIDS progression based upon survival analysis Cox proportional hazards, across *IL10* region of chromosome 1; (c) ARGRANK plots for HIV infection for SNPs across *CFTR* region of chromosome 7 for five ranking schemes; (d) ARGRANK plots for HIV/AIDS disease progression using case: control categories are the same 5 ranking schemes across chromosome 7; (e) ARGRANK plots for HIV infection across chromosome 3 including oSNPs, *CCR5-Δ32*, *CCR5-P1*, and *CCR2-64I*; (f) ARGRANK plots for HIV/AIDS disease progression 3 including oSNPs *CCR5-Δ32*, *CCR5-P1*, and *CCR2-64I*.

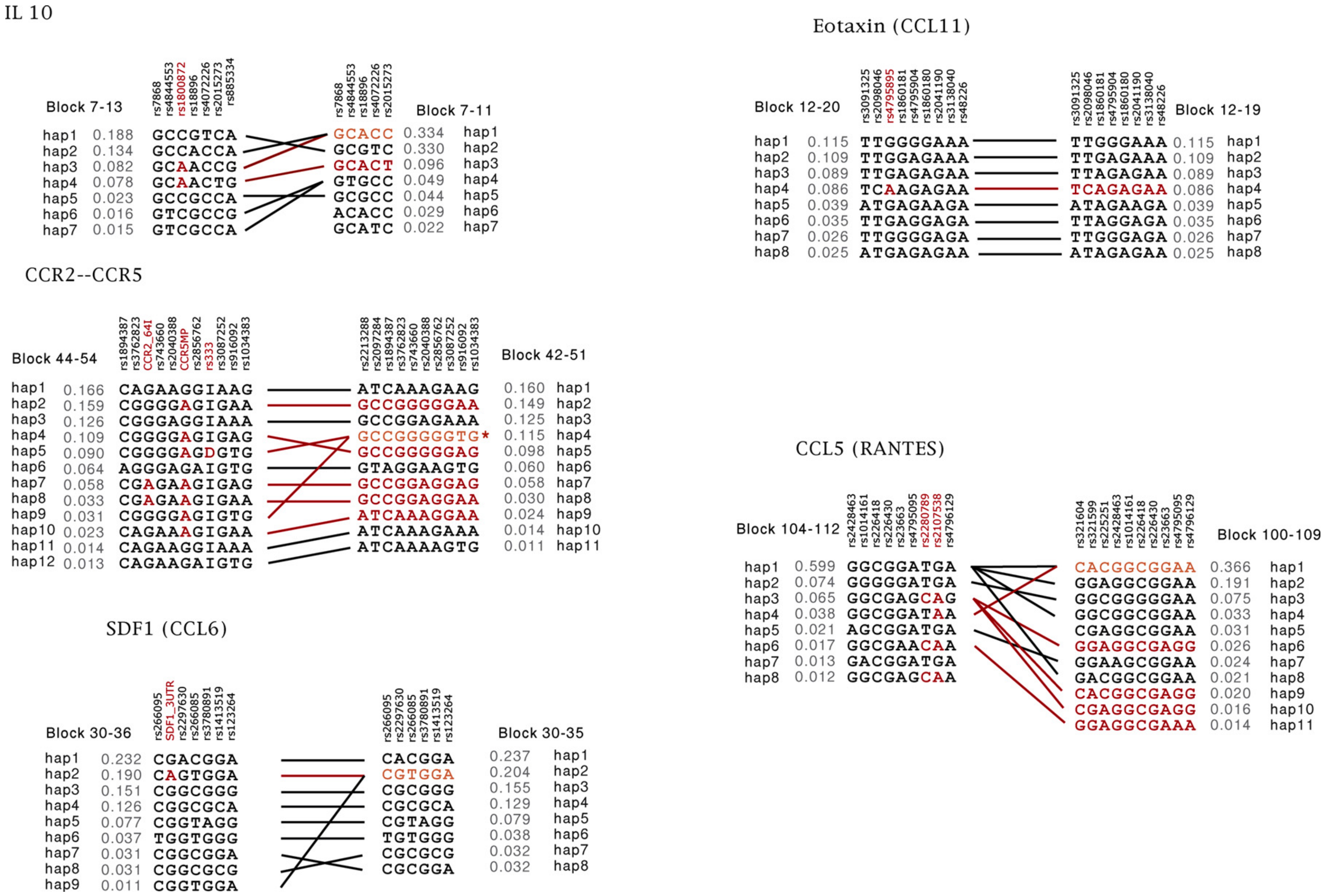

Figure 4 SNP composition of haplotypes imputed for five ARG regions studied here. For each region the left haplotype includes the oSNP and the right haplotypes overlap the oSNP locus but the oSNP is removed before haplotype imputation. Red lines illustrate the fate of oSNP containing haplotypes after the oSNP is removed. Frequency (*f*) of each haplotype is listed for all haplotypes. Percent SNP representation (PSR, see text) for the oSNP containing haplotypes is indicated in Table 5.

Table 5 Counts of significant tests using ARGARRAY and ARGRANK for oSNP, pSNP, haplotypes including oSNP and haplotypes built after excluding the oSNP

ARG	# Pats	oSNP		pSNP		oSNP detected[a]	Haplotype + oSNP							Haplotype-oSNP							
		Array #$P<.01$	Rank #$R<50$	Array #$P<.01$	Rank # $R<50$		Block	Hap	Freq	PSR (%)[b]	Array #$p<0.01$	Rank $>3R<160$	oSNP detected[a]	Block	Hap	Freq	Dilution	PSR (%)[b]	Array # $p<0.01$	Rank $>3R<160$	oSNP detected[a]
IL10	1490	11	15	12	3	Yes	7–13	H3	0.082	51%	0	0	No	7–11	H1	0.334	0.38	57%	0	0	Yes
								H4	0.078	49%	0	0			H3	0.096	1.00	43%	13	2	
CCR2-64I	1999	2	3	166	16	Yes	44–54	H7	0.058	64%	7	2	Yes	42–51	H7	0.058	1.00	66%	4	1	No
								H8	0.033	36%	0	0			H8	0.030	1.00	34%	0	0	
CCR5-P1	1953	2	5	46	4	Yes	44–54	H2	0.159	49%	3	2	Yes	42–51	H2	0.149	1.00	50%	10	3	Yes
								H4	0.109	34%	21	3			H4	0.115	0.26	10%	55	4	
								H9	0.031	10%	0	0			H5	0.098	1.00	33%	18	3	
								H10	0.023	7%	0	0			H9	0.024	1.00	8%	0	0	
CCR5-Δ32	2007	40	15	271	38	Yes	44–54	H5	0.09	100%	49	3	Yes	42–51	H4	0.115	0.74	100%	55	4	Yes
SDF1-3′A	2010	4	9	1	7	Yes	30–36	H2	0.19	100%	5	1	Yes	30–35	H2	0.204	0.95	100%	8	2	±
EOTAXIN	1961	0	2	7	4	±	12–20	H4	0.086	100%	0	0	No	12–19	H7	0.086	1.00	100%	0	0	No
RANTES-In1.1c	2005	0	0	0	0	No	104–112	H3	0.065	69%	0	0	No	100–109	H6	0.026	1.00	34%	16	1	No
								H6	0.017		0	0			H9	0.020	1.00	26%	0	0	
								H8	0.012	18%	2	0			H10	0.016	1.00	21%	0	0	
										13%					H11	0.014	1.00	18%	0	0	
RANTES-403	1920	0	0	0	0	No	104–112	H3	0.065	49%	0	0	No	100–109	H1	0.366	0.10	33%	1	1	No
								H4	0.038	29%	0	0			H6	0.026	1.00	23%	16	1	
								H6	0.017	13%	0	0			H9	0.020	1.00	18%	0	0	
								H8	0.012	9%	2				H10	0.016	1.00	14%	0	0	
															H11	0.014	1.00	12%	0	0	

[a] YES-signal apparent above background in counts ARGARRAY or ARGRANK or both, relative to ''negative control'' region background counts.

[b] PSR-percent SNP representation-see text and Fig. 4.

excluded contributed to weakened signal for the ARGs studied here.

How well did the GWA approach detect known ARGs?

The performance of SNPs and haplotypes in the ARG regions was compared by enumerating highly significant [$p<0.01$] signals revealed by ARGARRAY for each of the 136 tests (i.e. for four genetic hypotheses) for oSNPs, pSNPs, and haplotypes (Tables 4 and 5) and by determining the percentage of positive signals for different ARG oSNPs, pSNPs and haplotypes (Fig. 5). Then for the same ARG regions the number of times each oSNP, pSNP or haplotype showed a replicated low ARGRANK was assessed by how often a single SNP or haplotype had three or more ranks of less than 50 (in >3 of 5 ranking schemes as illustrated in Fig. 3a and b) for each of the four genetic hypothesis groups, (HIV-infection, categorical AIDS progression, survival rate, and sequelae). In Table 4, we list the counts for $R<50$ for the oSNP (left), but counts for $>3R<50$ for the pSNPs (Table 2, right) and the same level ($>3R<160$) for haplotypes with and without the oSNPs in Table 5. The gene discovery success for each ARG region was measured by comparing the percentage of significant replicate signals achieved oSNPs, pSNPs and haplotypes with ARGARRAY and ARGRANK (Fig. 5 and Table 5).

Our findings (Fig. 5, Table 5) suggest that for AIDS, a complex disease with multiple clinical stages to investigate, the computational tools do rather well is detecting oSNPs. Five ARGs (*CCR5-Δ32*, *CCR-5P1*, *CCR2-64I*, *IL10-5′A* and *SDF1-3′A*) gave strong statistical signals that suggest we would have discovered them with a pSNP approach. *EOTAXIN* pSNPs gave strong ARGRANK signals but produced low overall percentages, resulting in a ± equivocal call (Table 5). Given that *RANTES-405* and *RANTES-In1.1c* were not detected with oSNP analyses (Fig. 5a and b; Table 4) due to different *RANTES* haplotypes that carry offsetting ARGs influences [43–46], it is not surprising that these ARGs were not detected using these methods as well. The strongest signals occurred in the *CCR5/2* region (Figs. 2c and 3e, f) but associations were seen for other ARGs as well (Tables 4 and 5, Fig. 5). Both ARGARRAY and ARGRANK provided useful but complementary approaches for viewing large amounts of genetic association data, an important aspect in cases of large multi-variate cohort studies.

If the oSNPs are unknown, the search using pSNPs and haplotypes with and without the oSNP becomes important. The pSNPs performed consistently well while haplotype assessment varied considerably depending upon oSNP frequency, haplotype structure, haplotype frequency, PSR, haplotype re-structure upon removal of the oSNP (Fig. 5), as well as the occurrence of multiple associated signals in a region (e.g. as occurs within the *CCR5/2*; Figs. 2c and 3c). Twelve of the 24 ARG tests (Yes/No calls in Table 5) were detectable by pSNP or haplotypes providing an estimate that minimally 50% of the oSNP signals would have been discovered by pSNPs alone. As illustrated in Fig. 4, many if not most oSNPs are not carried on a unique haplotype (e.g. *SDF1 and EOTAXIN-CCL11* are the only ARGs carried on a unique haplotype among the SNPs genotyped here; Fig. 4) resulting in a dilution of the oSNP association signal when a single

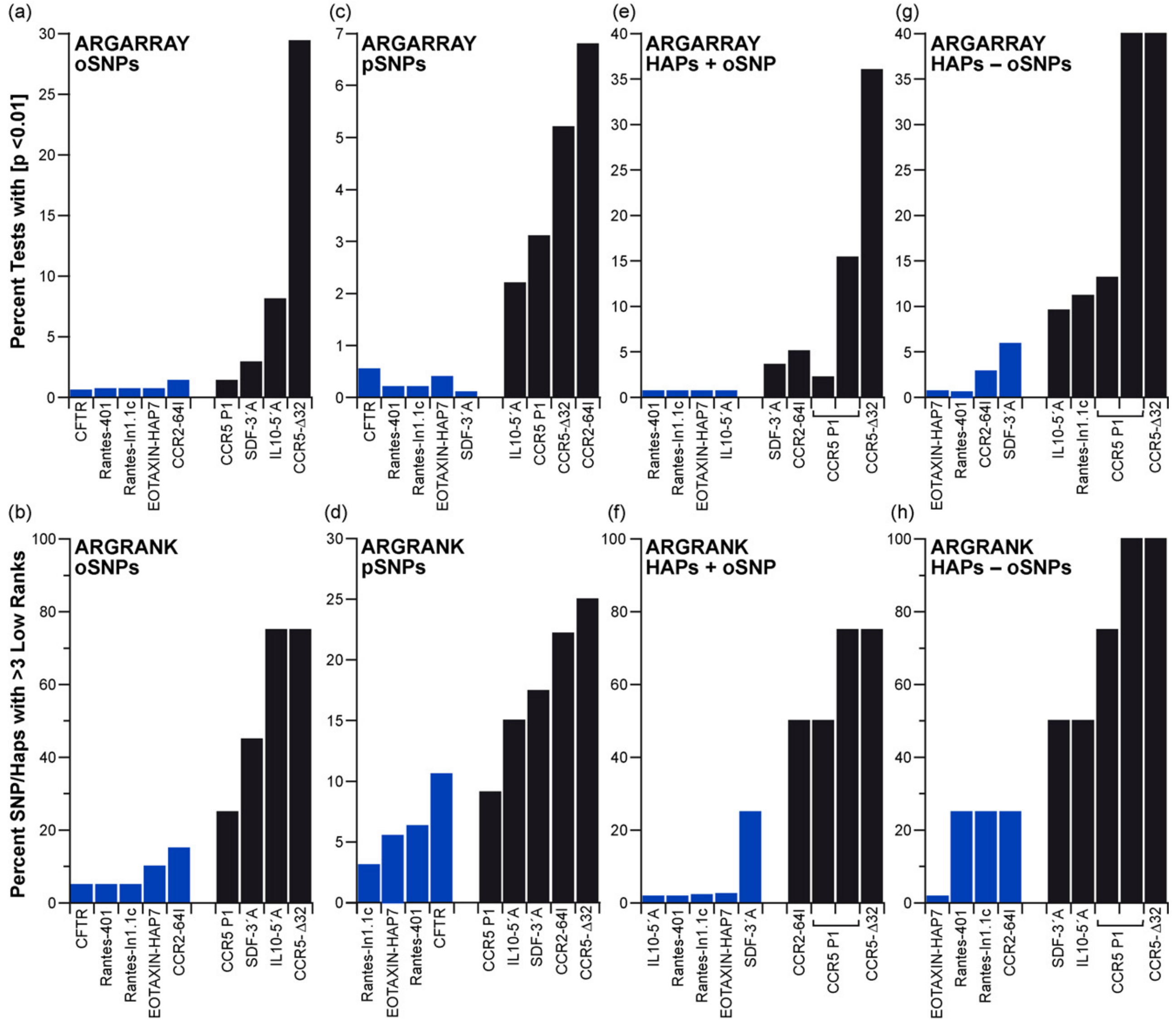

Figure 5 Percentages of significant positive genotype association tests for oSNPS (a and b), pSNPs (c and d), haplotypes bearing to oSNP (e and f), and haplotypes overlapping the oSNP site, but not included the oSNP allele in building the haplotypes (g and h). Percentages for ARGARRAY are the number of $p < 0.01$ tests/total tests run for oSNPs, pSNPs or haplotypes. For ARGRANK-oSNP, percentages equal the number of ranking schemes where the oSNP rank is < 50 out of 20 possible ranking schema (4 genetic hypotheses × 5 ranking schemes as described in Fig. 2 legend)/20. For ARGRANK-pSNP or haplotypes the percentages are the number times a pSNP ranks <50 in >3 of 5 schema/(5 ranking schemes × number of pSNPs or haplotypes). Raw counts for ARGARRAY and for ARGRANK are presented in Table 4.

haplotype carrying the oSNP is tested. This ''haplotype dilution'' effect reduces the strength of the genetic association signal and would produce false negatives. Perhaps a better advantage of haplotype definition lies in follow-up oSNP discovery within an associated chromosomal region. For such a region, saturated SNP genotyping can effectively narrow shared haplotypes' overlap among multiple individuals from an associated disease category, allowing one to close in upon the oSNP location more precisely.

By combining the results of pSNPs, haplotypes and algorithms for each ARG, 5—6 of the 8 ARGs studied (63—75%) were detected by pSNP or haplotype association, and a plausible explanation for the ARGs that failed can be offered. For example, within *RANTES* gene there occur three different AIDS restriction alleles (*In1.1C*, −*403A*, and −*28*) which produce offsetting influences on AIDS progression [44—46]. Interaction of these alleles was demonstrated in prior analyses and masks the effect in the present study as well (Supplemental Table 6). The previously reported *EOTAXIN-CCL11* influence on HIV infection [52] was also missed in our oSNP screen (Supplemental Table 6), although adjacent pSNPs did signal confirming that the original long haplotype association requires further haplotype dissection follow-up. The other ARGs selected did show signals and likely would have been discovered had they been unknown using the strategy described here.

Conclusions

Four principal conclusions can be drawn from our study. First, the results provide a useful transition from previous gene discoveries using single candidate gene variants to the high density GWA discovery in disease cohorts. Second, computational tools (BLOCKHEAD, ARGBROWSER, ARGARRAY, and ARGRANK) that render the challenge of multimillion-genotype/test datasets for complex disease gene detection feasible and tractable were evaluated empirically. The application of multiple tests about different genetic models and stages of AIDS pathogenesis adds a useful depth to our GWA screens by illustrating internal replication of SNPs that show a strong association signal. Third, this work illustrates the limits of haplotype-based GWA diminished by haplotype dilution of oSNPs. The strength of haplotype association seems to be more in closing in on the oSNP of an associated region than in detecting association signals in a disease cohort. Fourth, the pSNP approach works remarkably well in revealing oSNPs by capturing intrinsic LD around them. The oSNPs were detected by proxy almost as well as the oSNPs themselves and we project a minimum estimate for ARG discovery success as 50—75% of oSNPs with a blind genome scan of the scale described here (17 kb density, 2139 patients). These discoveries offer encouragement for the prospects of new ARG discoveries in the more dense 1000 K+ GWA design using the approaches described here as well as for other complex genetic diseases with multiple disease outcomes. The expectation of GWA studies now being undertaken in search of undiscovered ARGs is indeed promising.

Acknowledgements

Funding: This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.vaccine.2007.12.054.

References

[1] Cocchi F, De Vico AL, Garzino-Demo A, Arya SK, Gallo RC, Lusso P. Identification of RANTES, MIP-1α, and MIP-1β as the major HIV-suppressive factors produced by CD8 + T cells. Science 1995;270:1811—5.

[2] Dragic T, Litwin V, Allaway GP, Martin SR, Huang Y, Nagashima KA, et al. HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. Nature 1996;381:667—73.

[3] Alkhatib G, Combadiere C, Broder CC, Feng Y, Kennedy PE, Murphy PM, et al. CC CKR5: a RANTES, MIP-1α, MIP-1β receptor as a fusion cofactor for macrophage-tropic HIV-1. Science 1996;272:1955—8.

[4] Choe H, Farzan M, Sun Y, Sullivan N, Rollins B, Ponath PD, et al. The chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. Cell 1996;85:1135—48.

[5] Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, et al. Identification of a major co-receptor for primary isolates of HIV-1. Nature 1996;381:661—6.

[6] Doranz BJ, Rucker J, Yi Y, Smyth RJ, Samson M, Peiper SC, et al. A dual-tropic primary HIV-1 isolate that uses fusion and the β-chemokine receptors CKR-5, CKR-3, and CKR-2b as fusion cofactors. Cell 1996;85:1149—58.

[7] Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Science 1996;273:1856—62.

[8] Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, Horuk R, et al. Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. Cell 1996;86:367—77.

[9] Samson M, Labbe O, Mollereau C, Vassart G, Parmentier M. Molecular cloning and functional expression of a new human CC chemokine receptor gene. Biochemistry 1996;35:3362—7.

[10] O'Brien SJ, Nelson GW. Human genes that limit AIDS. Nat Genet 2004;36:565—74.

[11] Carrington M, O'Brien SJ. The influence of HLA genotype on AIDS. Annu Rev Med 2003;54:535—51.

[12] O'Brien SJ, Nelson GW, Winkler CA, Smith MW. Polygenic and multifactorial disease gene association in man: Lessons from AIDS. Annu Rev Genet 2000;34:563—91.

[13] Heeney JL, Dalgleish AG, Weiss RA. Origins of HIV and the evolution of resistance to AIDS. Science 2006;28:462—6.

[14] Manfredi R, Sabbatani S. A novel antiretroviral class (fusion inhibitors) in the management of HIV infection. Present features and future perspectives of enfuvirtide (T-20). Curr Med Chem 2006;13:2369—84.

[15] Aquaro S, D'Arrigo R, Scicher V, Di Perri G, Lo Caputo S, Visco-Comandini U, et al. Specific mutations in HIV-1 gp41 are associated with immunological success in HIV-1-infected patients receiving enfuvirtide treatment. J Antimicrob Chemother 2006;58:714—22.

[16] Lederman MM, Veazey RS, Offord R, Mosier DE, Dufour J, Mefford M, et al. Prevention of vaginal SHIV transmission in rhesus macaques through inhibition of CCR5. Science 2004;306:485—7.

[17] Bayes M, Rabasseda X, Prous JR. Gateways to clinical trial. Methods Find Exp Clin Pharmacol 2006;28:379—412.

[18] Sharmeen L, McQuade T, Heldsinger A, Gogliotti R, Domagala J, Gracheck S. CCR antagonists: host-targeted antivirals for the treatment of HIV infection. Antivir Chem Chemother 2005;16:339—54.

[19] Pereira CF, Paridaen JT. Anti-HIV drug development—an overview. Curr Pharm Des 2004;10:4005—37.

[20] De Clercq E. HIV-chemotherapy and prophylaxis: new drugs, leads and approaches. Int J Biochem Cell Biol 2004;36:1800—22.

[21] Nelson GW, O'Brien SJ. Using mutual information to measure the impact of multiple genetic factors on AIDS. J Acquir Immune Defic Syndr 2006;42:347—54.

[22] Nature Collections. Human Genome Supplement to Nature (June 1 edition); 2006. p. 1—305.

[23] The International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437:1299—320.

[24] International HapMap Consortium A second generation human haplotype map of over 3.1 million SNPs. Nature. (2007) Oct 18; 449851—61.

[25] Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. Mapping complex disease loci in whole-genome association studies. Nature 2004;429:446—52.

[26] Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 2006;38:1251—60.

[27] de Bakker PI, Burtt NP, Graham RR, Guiducci C, Yelensky R, et al. Transferability of tag SNPs in genetic association studies in multiple populations. Nat Genet 2006;38:1298—303.

[28] de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, et al. Efficiency and power in genetic association studies. Nat Genet 2005;37:1217—23.

[29] Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 2005;6:95—108.

[30] Risch NJ. Searching for genetic determinants in the new millennium. Nature 2000;405:847—56.

[31] Todd JA. Statistical false positive or true disease pathway? Nat Genet 2006;38:731—3.

[32] Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 2005;6:109—18.

[33] Clark AG, Boerwinkle E, Hixson J, Sing CF. Determinants of the success of whole-genome association testing. Genome Res 2005;15:1463—7.

[34] Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.

[35] Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet 2001;29:306—9.

[36] Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG. Genetic associations in large versus small studies: an empirical assessment. Lancet 2003;361:567—71.

[37] Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 2003;33:177—82.

[38] Arking DE, Pfeufer A, Post W, Kao WH, Newton-Cheh C, et al. A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. Nat Genet 2006;38:644—51.

[39] Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet 2007;39:207—11.

[40] Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 2007;316:1341—5.

[41] Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007;39:645—9.

[42] Welcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661—78.

[43] Duggal P, Winkler CA, An P, Yu XF, Farzadegan H, et al. The effect of RANTES chemokine genetic variants on early HIV-1 plasma RNA among African American injection drug users. J Acquir Immune Defic Syndr 2005;38(2005):584—9.

[44] An P, Nelson GW, Wang L, Donfield S, Goedert JJ, et al. Modulating influence on HIV/AIDS by interacting RANTES gene variants. Proc Natl Acad Sci USA 2002;99:10002—7.

[45] McDermott DH, Beecroft MJ, Kleeberger CA, et al. Chemokine RANTES promoter polymorphism affects risk of both HIV infection and disease progression in the Multicenter AIDS Cohort Study. AIDS 2000;14:2671—8.

[46] Gonzalez E, Dhanda R, Bamshad M, et al. Global survey of genetic variation in CCR5, RANTES, and MIP-1alpha: impact on the epidemiology of the HIV-1 pandemic. Proc Natl Acad Sci USA 2001;98:5199—204.

[47] Bashirova AA, Bleiber G, Qi Y, Hutcheson H, Yamashita T, Johnson RC, et al. TSG101 genetic variability shows consistent effects on multiple outcomes to HIV-1 exposure. J Virol 2006;80:6757—63.

[48] Martin MP, Lederman M, Hutcheson H, Nelson GW, Goedert JJ, Detels R, et al. Association of DC SIGN promoter polymorphisms with increased risk for parenteral but not mucosal acquisition of HIV-1 infection. J Virol 2004;78:14053—6.

[49] Javanbakht H, An P, Gold B, Petersen DC, O'hUigin C, Nelson GW, et al. Effects of human TRIM5a polymorphisms on antiretroviral function and susceptibility to human immunodeficiency virus infection. Virology 2006;354:15—27.

[50] An P, Duggal P, O'Brien SJ, Donfield S, Goedert JJ, Buchbinder S, et al. Polymorphisms of CUL5 are associated with CD4+ T cell loss in HIV-1 infected individuals. PLoS Genet 2007; 3:e19.

[51] An P, Wang LH, Hutcheson H, Nelson G, O'Brien SJ, Donfield S, et al. Functional regulatory polymorphisms in the gene encoding cyclophilin A influence HIV-1/AIDS. PLoS Pathogens 2007;3:849—57.

[52] Modi WS, Goedert JJ, Strathdee S, Buchbinder S, Detels R, Donfield S, et al. MCP-1-MCP-3-Eotaxin gene cluster influences HIV-1 transmission. AIDS 2003;17:2357—65.